View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Journal of Statistical Software

Journal of Statistical Software
April 2011, Volume 40, Issue 9. http://www.jstatsoft.org/

SPECIES: An R Package for Species Richness Estimation

Ji-Ping Wang
Northwestern University

Abstract

We introduce an R package **SPECIES** for species richness or diversity estimation. This package provides simple R functions to compute point and confidence interval estimates of species number from a few nonparametric and semi-parametric methods. For the methods based on nonparametric maximum likelihood estimation, the R functions are wrappers for Fortran codes for better efficiency. All functions in this package are illustrated using real data sets.

Keywords: coverage, Jackknife, nonparametric maximum likelihood estimation, R software, species problem, species richness estimation.

1. Introduction

1.1. Species richness estimation problem

The species problem has a wide range of important applications spanning multiple disciplines including ecology (Fisher *et al.* 1943; Boulinier *et al.* 1998), linguistics (Efron and Thisted 1976; McNeil 1973; Thisted and Efron 1987), numismatics (Stam 1987), and genomics (Mao 2002; Wang *et al.* 2005; Acinas *et al.* 2004; Hong *et al.* 2006). A typical species data set contains a series of counts $x_i, i = 1, ..., D$, recording the number of individuals observed from a total of D distinct species in the sample. The counts data are often further summarized into the *frequency of frequencies* data in the form of $\mathbf{n} = (n_1, ..., n_K)$ where $n_j = \sum_i^D I\{x_i = j\}$ (I is the indicator function) is the number of species with j individuals observed, and $K = \max_i x_i$ is the maximum number of individuals observed from any single species. In the following context, we shall reserve i for indexing the individual species, and j for the sample species abundance. The total number of the distinct species N in the underlying population is the parameter for estimation.

1.2. Overview of this package

A rich literature exists on this problem. For an excellent review, we recommend Bunge and Fitzpatrick (1993). There are a few well-known software tools available for computing species diversity, including **EstimateS** (Colwell 2009), **SPADE** (Chao 2010) and **ws2m** (Turner *et al.* 2003). These tools all offer a menu-driven interface to calculate the estimates for a single data set, but none provides the functionality that allows users to calculate the estimates repeatedly from the command line. This feature is appealing when one needs to systematically investigate the behavior of different estimators using Monte-Carlo simulations.

Recently, several new methods have been developed based on nonparametric maximum likelihood (NPML) estimation (Norris and Pollock 1998; Wang and Lindsay 2005; Wang 2010). These methods are competitive in performance while all complicated in computing. Therefore it is highly desirable to integrate these methods into a software tool. The R (R Development Core Team 2011) package **SPECIES** is a creation to this end, having seven main functions including `chao1984()`, `ChaoLee1992()`, `ChaoBunge()`, `jackknife()`, `unpmle()`, `pnpmle()` and `pcg()`, implementing the lower bound estimator by Chao (1984), two coverage-based estimators by Chao and Lee (1992), the coverage-duplication estimator by Chao and Bunge (2002), the Jackknife estimator by Burnham and Overton (1978, 1979), the unconditional NPML estimator (NPMLE) by Norris and Pollock (1998), the penalized conditional NPMLE by Wang and Lindsay (2005), and the the Poisson-Compound Gamma estimator by Wang (2010) respectively. The **SPECIES** package is available from the the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=SPECIES`.

2. Methods and package functions

All functions in **SPECIES** require the input data to be summarized in the format of *frequency of frequencies*. The input data, denoted as $\mathbf{n}$, must be defined as a two-column matrix or data frame, where the first column is j and the second column is n_j for $j = 1, ..., K$, sorted in the ascending order of j. The zero-frequencies ($n_j = 0$) can be omitted from $\mathbf{n}$. The following example is the famous Malayan butterfly data (Fisher *et al.* 1943) stored in **SPECIES**:

```
R> library("SPECIES")
R> data("butterfly")
R> butterfly

    j n_j
1   1 118
2   2  74
3   3  44
4   4  24
5   5  29
6   6  22
7   7  20
8   8  19
9   9  20
    :
25 25 119
```

In this study, a total of $D = 620$ distinct butterfly species were observed, of which, 118 were singletons. The frequency n_{25} denotes the collapsed count $\sum_{j \geq 25} n_j$. Other data sets stored in **SPECIES** include the expressed sequence tag(EST) data (`EST`, Wang *et al.* 2005), the microbial species data (`microbial`, Hong *et al.* 2006), the traffic data (`traffic`, Böhning and Schön 2005), cottontail rabbits data (`cottontail`, Chao 1987) and the insects data (`insects`, Burnham and Overton 1979).

The first three functions, `chao1984()`, `ChaoLee1992()`, and `ChaoBunge()`, implement multiple methods contributed by Chao and co-authors:

```
chao1984(n, conf = 0.95),

ChaoLee1992(n, t = 10, method = "all", conf = 0.95),

ChaoBunge(n, t = 10, conf = 0.95).
```

The argument `n` is the input data as described above. The argument `method` in `ChaoLee1992()` can be chosen as `ACE` or `ACE-1` (Chao and Lee 1992). One can also specify `method = "all"` (default) to compute both estimators. The argument `t` is an integer-valued cut-off that defines the less abundant ($j \leq t$) or more abundant species ($j > t$). The species data are often extremely right skewed. The less abundant species are more informative in predicting the number of the unseen species. The estimators `ACE`, `ACE-1`, and `ChaoBunge` are sensitive to the choice of t. Avoiding an over-large t helps reduce the risk of extreme variance or bias. The default value of `t` is `10` as suggested by the authors in their original papers. For the confidence interval with a specified level by the argument `conf` (default `0.95`), we used a log-transformation procedure from Chao (1987) for better coverage.

Let $D = \sum_{j=1}^{K} n_j$, and $T = \sum_{j=1}^{K} j \cdot n_j$. The `chao1984` estimator is a lower-bound estimator as follows:

$$\hat{N}_{\text{chao1984}} = D + 2n_1^2/n_2. \tag{1}$$

The estimator $\hat{N}_{\text{chao1984}}$ is simple, but typically biased downward as named (see a systematic investigation in Wang and Lindsay 2005, or the `traffic` data example below). The `ACE`, `ACE-1`, and `ChaoBunge` estimators are all based on a concept called *coverage*, denoted as C, defined in Good (1953) as follows:

$$C = \sum_{i=1}^{N} p_i I(x_i > 0), \tag{2}$$

where p_i is the relative abundance of species i, and $x_i = 0$ if species i is not observed. The term *coverage* measures the total abundance of observed species in the population. If the species abundance p_i is homogeneous across all species, then clearly $N = D/C$. An estimator of the C from Good (1953) is $\hat{C} = 1 - n_1/T$, resulting in the Good's estimator

$$\hat{N}_{\text{G}} = D/\hat{C}. \tag{3}$$

Like $\hat{N}_{\text{chao1984}}$, the Good's estimator $\hat{N}_G$ is well known for under-estimation because natural species populations are typically heterogeneous(Chao and Lee 1992). To account for the heterogeneity of species abundance, Chao and Lee (1992) proposed two improved estimators by estimating the coefficient of variation (CV) of p_i. The resulting estimators are

$$\hat{N}_{\text{ACE}} = \frac{D}{\hat{C}} + \frac{T(1-\hat{C})}{\hat{C}}\hat{\gamma}^2, \tag{4}$$

$$\hat{N}_{\text{ACE}-1} = \frac{D}{\hat{C}} + \frac{T(1-\hat{C})}{\hat{C}}\tilde{\gamma}^2, \tag{5}$$

where $\hat{\gamma}$ and $\tilde{\gamma}$ are two estimators of CV. In particular the second provides further bias correction beyond $\hat{\gamma}$, but typically incurring larger variance of $\hat{N}_{\text{ACE}-1}$.

For the `ChaoBunge` method, a Gamma mixed Poisson model was assumed. Let $\theta = \mathsf{P}(X \geq 2)$. An estimator of the duplication proportion $\hat{\theta}$ was proposed in Chao and Bunge (2002), giving the following extrapolation estimator

$$\hat{N}_{\text{ChaoBunge}} = \sum_{j=2}^{K} n_j/\hat{\theta}. \tag{6}$$

The fourth function `jackknife()` computes the jackknife estimator by Burnham and Overton (1978, 1979),

```
jackknife(n, k = 5, conf = 0.95).
```

The kth order jackknife estimator is

$$\hat{N}_{\text{Jk}} = D + \sum_{j=1}^{k}(-1)^{j+1}\binom{k}{j}n_j. \tag{7}$$

The Jackknife order is used to balance bias and variance. A higher order corrects for more bias, but causes larger variance as well. The Jackknife order is specified by the argument `k` with a default value `5`. This function also automatically computes the order using a step-wise testing procedure from the above papers. The argument `conf` specifies the confidence level not only for the confidence interval, but also for the critical value used for the step-wise Z test to determine the order. If the specified order is larger than the obtained from the test, then the latter is used in the output. If the test-selected order exceeds 10, the estimate at $k = 10$ will be reported (which though rarely happens in practice).

The rest three functions `unpmle()`, `pnpmle()` and `pcg()` are three variants of the nonparametric maximum likelihood based approaches under a Poisson mixture model. They were all wrapped from Fortran codes for computing speed consideration. Suppose X follows a Poisson mixture distribution $f(x;Q)$ where the mixing distribution is Q for the mean parameter. Then $\mathbf{n} = (n_1, ..., n_K\}$ follows a multinomial distribution with corresponding cell probabilities $f(j;Q), j = 1, ..., K$. The resulting likelihood can be factored into two parts as follows:

$$L(N, Q; \mathbf{n}) \propto \binom{N}{D} f(0;Q)^{N-D}\{1 - f(0;Q)\}^D \times \prod_{j>0}\left\{\frac{f(j;Q)}{1-f(0;Q)}\right\}^{n_j} := L^m \times L^c, \tag{8}$$

where L^m and L^c are referred to as the marginal and conditional likelihood respectively (conditioning on that a species is observed, e.g., for n_j's with $j > 0$).

The function `unpmle()` calculates the unconditional NPMLE of the species number from Norris and Pollock (1998),

```
unpmle(n, t = 15, C = 0, method = "W-L", b = 200, seed = NULL, conf = 0.95,
    dis = 1).
```

In this approach, a pair of $(\hat{N}, \hat{Q})$ is found to maximize the unconditional likelihood $L(N, Q; \mathbf{n})$. The argument `t` is the same cut-off as described in the `ChaoLee1992()` function with a default value `15`. We recommend to use $t \geq 10$. The argument `C` (= `1` or `0`) specifies whether a confidence interval should be calculated. Since there is no analytical form for the confidence interval, a bootstrap confidence interval of level specified by argument `conf` is provided (Wang 2010). The arguments `b` and `seed` specify how many bootstrap samples to be generated, and what seed to be used in random number generation for bootstrap respectively. If `seed` is not specified, the R internal random seed is used. These two arguments are ignored if `C = 0`. The argument `method` specifies which method to be used to find the unconditional NPMLE of N. The first method is `"N-P"`, in which an iterative algorithm by Böhning and Schön (2005) is used. Sometimes this method can be extremely slow. Alternatively one can use the default method `"W-L"` by Wang and Lindsay (2005), in which the approximate unconditional NPMLE (with high precision) of Q is found from the following penalized likelihood:

$$\log L \approx \log L^c - 0.5 \log\{f(0; Q)\}. \tag{9}$$

The approximate method typically yields identical or nearly identical estimate as the exact method, while it can be drastically faster (see illustrations below). The unconditional NPMLE can be as extreme as ∞. If the point estimate progresses beyond $20 \cdot D$ within iterations, the algorithm stops and reports the current point estimate. Otherwise the reported unconditional NPMLE can be even more unstable. The last argument `dis` specifies whether the mixture estimates should be output to the screen. Turning this off by setting `dis = 0` allows to avoid overflow of screen information in Monte-Carlo simulations.

To improve the stability of the NPML estimators, Wang and Lindsay (2005) proposed a penalized NPMLE by applying a quadratic penalty function to the conditional likelihood. This method is implemented in function `pnpmle()`,

```
pnpmle(n, t = 15, C = 0, b = 200, seed = NULL, conf = 0.95, dis = 1).
```

All the arguments are the same as described in `unpmle()`.

The last function `pcg()` calculates the Poisson-compound Gamma estimator by Wang (2010),

```
pcg(n, t = 35, C = 0, alpha = c(1:10), b = 200, seed = NULL, conf = 0.95,
    dis = 1).
```

This method was motivated by severe under-estimation observed from popular nonparametric estimators due to interplay of inadequate sampling effort, large heterogeneity and skewness (Wang and Lindsay 2005). Unlike `unpmle` or `pnpmle` method where the species abundance distribution is estimated by a discrete distribution, a compound Gamma with a unified shape parameter (α) is used in `pcg` method to bring more bias correction in targeted situations. The unified shape parameter α is chosen by a cross-validation procedure on a grid specified by the argument `alpha` to balance the bias and variance of the resulting estimates of zero-truncated Poisson mixture probabilities. This function automatically appends $\alpha = \infty$ onto the grid for cross-validation. We recommend to use a grid with $\alpha \geq 1$ to avoid extreme variability. The other arguments are the same as defined in `pnpmle()` or `unpmle()`. We also recommend to use a relatively larger `t` than the `unpmle()` or `pnpmle()` (default value is `35`) since we are fitting a continuous curve for Q. Caution should be taken if the last count in $\mathbf{n}$, n_K, is a

collapsed count of species that have $x \geq K$. For example, in the butterfly data from this package, n_{25} stands for $\sum_{j \geq 25} n_j$, and therefore $t \leq 24$ should be used.

3. Illustrations

In this section, we illustrate all main functions using the data sets from original publications. If the result differs from the reported, it is explicitly pointed out below. We first illustrate the `chao1984()` function using the `cottontail` rabbit data from Chao (1987, p. 787). This data set was from a capture-recapture experiment. The species number estimation methods also apply to this type of data.

```
R> library("SPECIES")
R> data("cottontail")
R> cottontail

  j n_j
1 1  43
2 2  16
3 3   8
4 4   6
5 5   0
6 6   2
7 7   1

R> chao1984(cottontail)

$Nhat
[1] 134

$SE
[1] 24.02129

$CI
      lb  ub
[1,] 102 202
```

The reported point estimate and 95% confidence interval in Chao (1987) were 134 and (103,202) respectively. The minor difference in the lower bound is probably due to rounding error. As another illustration, we applied `chao1984()` to the Taxicab data from Sampling scheme B.g in Table 1 of Chao (1987), the results are identical.

```
R> n = cbind(c(1:4), c(116, 48, 6, 2))
R> chao1984(n)

$Nhat
[1] 312
```

```
$SE
[1] 35.02778

$CI
      lb  ub
[1,] 259 399
```

For `jackknife()` function, we illustrate it using the `insects` data analyzed in Burnham and Overton (1979, p. 935) at $k = 2$ (note: n_6 below is the collapsed count for $j \geq 6$).

```
R> data("insects")
R> insects

  j n_j
1 1  50
2 2  20
3 3  11
4 4   6
5 5   5
6 6  32

R> jackknife(insects, k = 2)

$JackknifeOrder
[1] 2

$Nhat
[1] 204

$SE
[1] 17.32051

$CI
      lb  ub
[1,] 170 238
```

To further verify the authenticity of the function, I reproduced the Table 6 of Burnham and Overton (1979, p. 935). The results are presented in Table 1.

Note the selected order based on the stepwise test is 2 at significance level 0.05. Therefore if the user had specified a higher order, the output would still be the same, e.g., at order = 2. We noticed that results in Table 1 are identical to the original results (with negligible rounding errors) except for T_k and P_k at $k = 1$. Further work is need to figure out the cause of this slight but apparent discrepancy. So far the author has not yet found any discrepancy between the output from this function and the published reports in terms of point estimate and standard error.

k	$\hat{N}_{Jk}$	$se(\hat{N}_{Jk})$	T_k	P_k
1	174	10.00	3.772	0.00016
2	204	17.32	1.784	0.0744
3	225	27.23	0.928	0.353
4	242	42.66	0.576	0.565
5	259	68.12	-	-

Table 1: Reproduction of Table 6 of Burnham and Overton (1979, p. 935). The order k jackknife estimate is denoted as $\hat{N}_{Jk}$, and $se(\hat{N}_{Jk})$ is its standard error . The test statistic T_k is the Z-statistic and P_k is the two-sided p-value for testing order= k vs. order= $k+1$.

We illustrate `ChaoLee1992()`, `ChaoBunge()`, `unpmle()`, and `pnpmle()` using the `butterfly` data that was analyzed in Chao and Bunge (2002, Table 2, p. 535) and Wang and Lindsay (2005, Table 2, p. 949).

```
R> data("butterfly")
R> ChaoLee1992(butterfly, t = 10, method = "all")

$Nhat
[1] 712 737

$SE
[1] 17.35141 23.93183

$CI
       lb  ub
ACE   680 748
ACE-1 693 787

R> ChaoBunge(butterfly, t = 10)

$Nhat
[1] 757

$SE
[1] 32.39362

$CI
      lb  ub
[1,] 698 826

R> unpmle(butterfly, t = 15, method = "N-P")

Method: Unconditional NPMLE method by Norris and Pollock 1996, 1998,
        using algorithm by Bonhing and Schon 2005:
```

```
        MLE=                                722
        Estimated Poisson mixture components:
        p=                                  1.107039 4.370269 9.584652
        pi=                                 0.5693337 0.1878771 0.2427893

$Nhat
[1] 722

R> unpmle(butterfly, t = 15, method = "W-L", C = 1)

Method: Unconditional NPMLE method by Norris and Pollock 1996, 1998,
        using algorithm by Wang and Lindsay 2005:

        MLE=                                722
        Estimated Poisson mixture components:
        p=                                  1.110267 4.378968 9.586036
        pi=                                 0.5696068 0.1875711 0.242822

Start bootstrap 200  times:
......

$Nhat
[1] 722

$CI
      lb  ub
[1,] 688 920

R> pnpmle(butterfly, t = 15, C = 1)

Method: Penalized NPMLE method by  Wang and Lindsay 2005.

        MLE=                                724
        Estimated zero-truncated Poisson mixture components:
        p=                                  1.090829 4.326313 9.57749
        pi=                                 0.4675136 0.2317262 0.3007601

Start bootstrap  200  times:
......

$Nhat
[1] 724

$CI95
      lb  ub
[1,] 690 858
```

For `unpmle()`, it took about 4 minutes to compute the bootstrap confidence interval based on 200 samples using the approximate method `"W-L"` on a Mac OSX machine with a 2.93 GHz processor. The exact method `"N-P"` uses an algorithm by Böhning and Schön (2005), treating the unobserved species as missing data. Its computing time depends on the fraction of the missing information. For example, in the butterfly data, about 14% of the species were not observed (based on the point estimate $\hat{N} = 722$). It took the exact method about 20 minutes to finish 200 bootstrap estimates. In the traffic data below, about 83% of the species were not observed. As a result it took about 2 hours to finish 200 bootstrap samples using the exact method in contrast to about 4 minutes using the `"W-L"` method.

The `pnpmle` and `pcg` methods are both based on the conditional likelihood. The NPML estimates (`p` and `pi`) from `pnpmle()` are the mean parameters and their respective weights in the zero-truncated Poisson mixture. For `pcg()`, the output `p` and `pi` are the mean parameters and their weights of the Gamma mixture in the zero-truncated Poisson-compound Gamma model under the selected α model. The computing time for `pnpmle()` is typically a few minutes based on 200 bootstrap samples. The `pcg()` procedure is much more computing intensive because of a cross-validation procedure used in model selection. It is common to take more than one hour for 200 bootstrap samples. We illustrate it with the traffic data, which originally appeared in Simar (1976) and was reanalyzed recently by Böhning and Schön (2005) and Wang (2010). Since the true N is known as 9461, we include all other estimators for a comparison.

```
R> data("traffic")
R> traffic
```

```
  j  n_j
1 1 1317
2 2  239
3 3   42
4 4   14
5 5    4
6 6    4
7 7    1
```

```
R> chao1984(traffic)
```

```
$Nhat
[1] 5250

$SE
[1] 314.1841

$CI
       lb   ub
[1,] 4684 5919
```

```
R> ChaoLee1992(traffic, t = 7)
```

```
$Nhat
[1] 5684 6788

$SE
[1] 363.7709 648.4647

$CI
        lb   ub
ACE   5031 6461
ACE-1 5665 8223

R> ChaoBunge(traffic, t = 7)

$Nhat
[1] -21023

$SE
[1] 29020.49

$CI
        lb      ub
[1,] -1659 -154706

R> jackknife(traffic)

$JackknifeOrder
[1] 5

$Nhat
[1] 6170

$SE
[1] 256.7645

$CI
       lb   ub
[1,] 5667 6673

R> Good = sum(traffic[, 2])/(1 - traffic[1, 2]/sum(traffic[, 1] *
+     traffic[, 2]))
R> Good

[1] 4623.612

R> unpmle(traffic, t = 7, C = 1)
```

```
Method: Unconditional NPMLE method by Norris and Pollock 1996, 1998,
        using algorithm by Wang and Lindsay 2005:

        MLE=                                5497
        Estimated Poisson mixture components:
        p=                                  0.3360416 2.549361
        pi=                                 0.9851394 0.01486064


Start bootstrap 200  times:
......

$Nhat
[1] 5497

$CI
       lb    ub
[1,] 4948 17899
```

R> pnpmle(traffic, t = 7, C = 1)

```
Method: Penalized NPMLE method by  Wang and Lindsay 2005.

        MLE=                                5496
        Estimated zero-truncated Poisson mixture components:
        p=                                  0.3360911 2.549769
        pi=                                 0.9535588 0.04644124

Start bootstrap  200  times:
......


$Nhat
[1] 5496

$CI
       lb   ub
[1,] 4887 6895
```

R> pcg(traffic, C = 1, t = 35)

```
Method: Poisson-Compound Gamma method by Wang 2010.
Alpha grid used: 1 2 3 4 5 6 7 8 9 10 .

        MLE=                                6935
        Selected alpha model:               3
        Estimated Gamma components:
```

```
        p=                                  0.2624849 1.600614
        pi=                                 0.9304468 0.06955318

Start bootstrap  200  times:
......

$Nhat
[1] 6935

$AlphaModel
[1] 3

$CI95
[1]  5059 13396
```

Clearly N is substantially under-estimated by most of the estimators except `unpmle` and `pcg`. The negative estimate was observed for `ChaoBunge` method probably because the true species abundance distribution deviated from the Gamma distribution assumed in this method (see also Wang and Lindsay 2005 and Wang 2010).

4. Discussion

Many methods exist on the species problem. Most of the methods included in this package feature robust behavior regardless of the true form of the species abundance distribution Q. However, complexity in calculation poses challenges to ordinary users. For example, although the `ACE` and `ACE-1` estimators included in `ChaoLee1992()` both have analytical forms, their standard error calculation can be very complicated. Likewise, computing the NPML estimators can be intimidating to users that are unfamiliar with NPML estimation. This package is hoped to facilitate the dissemination of these methods to ordinary users in statistics and other disciplines.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004). "Fine-scale Phylogenetic Architecture of a Complex Bacterial Community." *Nature*, **430**, 551–554.

Böhning D, Schön D (2005). "Nonparametric Maximum Likelihood Estimation of Population Size Based on the Counting Distribution." *Journal of the Royal Statistical Society C*, **54**, 721–737.

Boulinier T, Nichols JD, Sauer JR, Hines JE, Pollock KH (1998). "Estimating Species Richness: The Importance of Heterogeneity in Species Detectability." *Ecology*, **79**(3), 1018–1028.

Bunge J, Fitzpatrick M (1993). "Estimating the Number of Species: A Review." *Journal of the American Statistical Association*, **88**(421), 364–373.

Burnham KP, Overton WS (1978). "Estimation of the Size of a Closed Population When Capture Probabilities Vary among Animals." *Biometrika*, **65**(3), 625–633.

Burnham KP, Overton WS (1979). "Robust Estimation of Population Size When Capture Probabilities Vary among Animals." *Ecology*, **60**(5), 927–936.

Chao A (1984). "Nonparametric Estimation of the Number of Classes in a Population." *Scandinavian Journal of Statistics, Theory and Applications*, **11**(4), 265–270.

Chao A (1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics*, **43**(4), 783–791.

Chao A (2010). ***SPADE**: Species Prediction and Diversity Estimation*. URL http://chao.stat.nthu.edu.tw/softwareCE.html.

Chao A, Bunge J (2002). "Estimating the Number of Species in a Stochastic Abundance Model." *Biometrics*, **58**(3), 531–539.

Chao A, Lee SM (1992). "Estimating the Number of Classes via Sample Coverage." *Journal of the American Statistical Association*, **87**(417), 210–217.

Colwell RK (2009). ***EstimateS**: Statistical Estimation of Species Richness and Shared Species from Samples*. Version 8.2, URL http://viceroy.eeb.uconn.edu/estimates.

Efron B, Thisted R (1976). "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?" *Biometrika*, **63**(3), 435–447.

Fisher RA, Corbet AS, Williams CB (1943). "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population." *Journal of Animal Ecology*, **12**(1), 42–58.

Good IJ (1953). "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika*, **40**(3/4), 237–264.

Hong SH, Bunge J, Jeon SO, Epstein SS (2006). "Predicting Microbial Species Richness." *Proc. Natl. Acad. Sci.*, **103**(1), 117–122.

Mao C (2002). *Mixture Models for Species and Population Size Estimation*. Ph.D. thesis, The Pennsylvania State University.

McNeil D (1973). "Estimating an Author's Vocabulary." *Journal of the American Statistical Association*, **68**(341), 92–96.

Norris JL, Pollock KH (1998). "Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity between Species." *Environmental and Ecological Statistics*, **5**(4), 391–402.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistcial Computing. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Simar L (1976). "Maximum Likelihood Estimation of a Compound Poisson Process." *The Annals of Statistics*, **4**(6), 1200–1209.

Stam AJ (1987). "Statistical Problem in Ancient Numismatics." *Statistica Neerlandica*, **41**(3), 151–173.

Thisted R, Efron B (1987). "Did Shakespeare Write a Newly-Discovered Poem?" *Biometrika*, **74**(3), 445–455.

Turner W, Leitner W, Rosenzweig M (2003). ***ws2m**: Software for the Measurement and Analysis of Species Diversity*. URL http://eebweb.arizona.edu/diversity.

Wang JP (2010). "Estimating the Species Richness by a Poisson-Compound Gamma Model." *Biometrika*, **97**(3), 727–740.

Wang JPZ, Lindsay BG (2005). "A Penalized Nonparametric Maximum Likelihood Approach to Species Richness Estimation." *Journal of American Statistical Association*, **100**(471), 942–959.

Wang JPZ, Lindsay BG, Cui L, Wall PK, Marion J, Zhang J, dePamphilis CW (2005). "Gene Capture Prediction and Overlap Estimation in EST Sequencing from One or Multiple Libraries." *BMC Bioinformatics*, **6**(300).

Affiliation:

Ji-Ping Wang
Department of Statistics
Northwestern University
2006 Sheridan Road
Evanston, IL 60208, United States of America
E-mail: jzwang@northwestern.edu
URL: http://bioinfo.stats.northwestern.edu/~jzwang/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 40, Issue 9 *Submitted:* 2010-03-11
April 2011 *Accepted:* 2011-03-27